5-2017

Identification of Dynamic Outliers

Kangkana Sarmah Baruah

Title of Thesis: IDENTIFICATION OF DYNAMIC OUTLIERS

Kangkana Sarmah Baruah, Master of Science, 2017

Thesis directed by: Dr. Andrada E. Ivanescu

Department of Mathematical Sciences

Abstract

Several methods for performing the identification of outliers are described when dealing with functional data. The methods studied include prediction intervals for detection of dynamic functional outliers as well as related methods from the functional data literature. A comparison of methods is performed using metrics for dynamic outlier identification. Simulations and applications to environmental studies illustrate the applicability of the methods. Results obtained from simulation and application to real dataset suggest that Dynamic Function-on-Function Regression is a preferable method for detecting dynamic outliers. This method can detect outliers at a very high identification rate. Identification rate of dynamic outliers increases when a large number of curves and large size of outliers is observed.

MONTCLAIR STATE UNIVERSITY

IDENTIFICATION OF DYNAMIC OUTLIERS

by

Kangkana Sarmah Baruah

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Statistics – Master of Science

May 2017

College/School College of Sciences and Mathematics

Department Mathematical Sciences

Thesis Committee:

[redacted]

Dr. Andrada E. Ivanescu

Thesis Sponsor

[redacted]

Dr. Haiyan Su

Committee Member

[redacted]

Dr. Andrew McDougall

Committee Member

IDENTIFICATION OF DYNAMIC OUTLIERS

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science in Statistics

by

KANGKANA SARMAH BARUAH

Montclair State University

Montclair, NJ

2017

Contents

List of figures

List of tables

Table Page No.

Identification of Dynamic Outliers

1 Introduction

Functional data are high dimensional data having a rich source of information which brings many opportunities for research and data analysis. With the furtherance of technology, data are being recorded along a continuum over some domain, such as a time interval, at different discrete points, such as time points. Since observed data can be sampled over location, time, and some other domain, these data can be considered as observed discretized functions on some domain. Although these data are higher dimensional objects, the functional data can be collected and stored in a finite dimensional form typically on some fine grid. These types of data are known as functional data which are considered as realizations of random functions collected at discrete points in the domain. Functional data are modeled as samples of smooth random trajectories with additive noise. Functional Data Analysis (FDA) was coined by Ramsay (1982), Ramsay and Dalzell (1991) and Ramsay and Silverman (2005). This type of data is increasingly being used for better analysis, modeling, and prediction. Popular topics in FDA include the study of the prediction of functional data, assessing relationships between functional random variables and other quantities which cannot be analyzed by traditional methods.

The analysis of functional data has its importance in different scientific fields of interest, such as medicine (Sorensen et al., 2013), environmental sciences (Caballero et al.,

2013; Martinez et al., 2011), and monitoring weather and air quality (Ignaccolo et al., 2013). Some important facets of FDA comprise the choice of smoothing technique, reduction of data using functional principal components, functional linear models, and forecasting. Common goals of FDA are to start with some exploratory analysis, to represent and display data in order to highlight characteristics of interest, and then to continue and use them for further analysis. Specific statistical methodology for FDA include the characterization of functional mean (Bunea et al., 2011), function-on-function regression (Ivanescu et al., 2015), estimations of individual curves from noisy observed data (Goldsmith et al., 2013), characterizing the patterns of variability among curves (Ramsay and Silverman 2005; Yao et al., 2005), prediction using functional regression model (Goldsmith and Scheipl, 2014). Functional prediction is one of the important aspects of FDA. Dynamic prediction for functional data has been investigated by Chiou (2012), Shang (2015), Goldberg et al. (2014). Dynamic predictions are made on the basis of available data up to some time points.

An important aspect of FDA is to detect outliers. Outliers are observed functional points far away from other functional observations. Although some research has been carried out to detect outliers in regression problems with univariate and multivariate sample data, some recent work has also been done in FDA. In our work we rely on a dynamic approach for dynamic outlier detection in functional data. Several measures for identification are used. We detect outliers for the future trajectory on the basis of observed data until a specific time point. Although tremendous work has been done in different areas of the FDA literature, detecting dynamic functional outliers is novel for functional data analysis. In the area of prediction with functional data, methods were put forward by Chiou (2012), where dynamic prediction is discussed. In the area of outlier detection Sawant et

al. (2012) discusses functional outlier detection. Functional data analysis includes several other methods for outlier detection using measures of function depth (Febrero et al., 2007; Febrero et al., 2008; Febrero et al., 2012).

2 Research Method

In this work our principal goal is to introduce and study the performance of Dynamic Penalized Function-on-Function Regression (DPFFR) method for detecting dynamic outliers. We are using DPFFR to obtain dynamic functional predictions. By approximating the integrals via Riemann sums on a dense grid for domain $t \in T$, where $T \in \{1,2,\dots,r\}$, n functional responses $Y_i(\tilde{t})$, $1 \leq i \leq n$, over time domain $\tilde{t} \in \{r+1, r+2, \dots, M\}$, are assumed to be expressed by the model:

$$Y_i(\tilde{t}) = W_{i1}.\gamma_1 + W_{i2}.\gamma_2 + \zeta(\tilde{t}) + \int_T Y_i(t)\beta(\tilde{t},t)dt + \int_T Z_i(t)\delta(\tilde{t},t)dt + \epsilon_i(\tilde{t}). \quad (1)$$

The curves $Y_i(t)$ were historical curves observed on an equally spaced grid $t \in \{1, 2, \dots, r\}$ and the functional covariate $Z_i(t)$ was observed at the same equally spaced grid $t \in \{1,2,\dots,r\}$. Model parameters $\beta(\tilde{t},t)$ and $\delta(\tilde{t},t)$ are bivariate functional parameters, $\zeta(\tilde{t})$ is the functional intercept, $\epsilon_i(\tilde{t})$ are random errors, W_{i1} and W_{i2} are scalar covariates. For higher values of the bivariate parameters $\beta(\tilde{t},t)$ and $\delta(\tilde{t},t)$, we would observe more influence on the functional responses $Y_i(\tilde{t})$.

For example, let us consider $Y_i(\tilde{t})$ and $Y_i(t)$ be the outdoor humidity for the second and first half of the day respectively where measurements are taken every hour. Predictor $Z_i(t)$ could be the temperature for the first half of the day which is also measured hourly. Model parameter $\beta(\tilde{t},t)$ would be greater when the humidity in the first half is

useful for the prediction for the second half. Similarly, $\delta(\tilde{t}, t)$ would be greater when the temperature in the first half of the day is useful for predicting humidity in second half of the day.

Since functional data are noisy we want to apply smoothing by eliminating roughness while retaining the right shape when model parameters are estimated. Penalized regression ensures smoothness for resulting functional parameters of the model. We use dynamic penalized function-on-function regression (DPFFR) for estimation of parameters, dynamic prediction, prediction intervals and identification of dynamic outliers. For the model parameters a large number of basis functions are chosen and penalties $\lambda_\zeta P(\zeta), \lambda_\beta P(\beta),$ and $\lambda_\delta P(\delta)$ are applied. If we denote by $f_{i,\tilde{t}}(\gamma_1, \gamma_2, \zeta, \beta, \delta)$ the mean of $Y_i(\tilde{t})$, the penalized criterion to be minimized is

$$\sum_{i,\tilde{t}} ||Y_i\{\tilde{t}\} - f_{i,\tilde{t}}(\gamma_1, \gamma_2, \zeta, \beta, \delta)||^2 + \lambda_\zeta P(\zeta) + \lambda_\beta P(\beta) + \lambda_\delta P(\delta),$$

which is a penalized least squares criterion and corresponds to the residual sum of squares criterion specific for the setting of dynamic penalized functional regression. With the estimated model parameters dynamic prediction is performed, and then dynamic outliers are detected based on the prediction interval. In dynamic prediction we predict the future values on the basis of available observations until some specific point r.

We use several metrics for identification of dynamic outliers. We calculate the rate of detection of outliers as a frequency that is calculated: i) as the average frequency across all grid points where predictions are made, and ii) as a function of $\tilde{t} \in \{r+1, r+2, \ldots, M\}$. Based on the prediction intervals we also compute : 1) the width of the confidence intervals, and 2) the distance between the dynamic outlying observation and the closest upper/lower bound of the prediction. These measures are displayed : i) as the average

across all grid points where predictions are made, and ii) as a function of $\tilde{t} \in \{r+1, r+2, \ldots, M\}$.

3 Comparison of Methods for Identification of Dynamic Outliers

Different methods have been developed for identifying dynamic outliers. In our work the methods BENDY, DLM, DPFR, dynamic_FLR and dynupdate are considered. These methods are compared using several metrics for dynamic outlier identification. For each method, the prediction interval for the future trajectory is obtained. The observations that are positioned outside the prediction intervals are considered dynamic outliers.

3.1 BENDY

BENchmark DYnamic model (BENDY) is used for predicting the response for a particular time domain ($\tilde{t}$). The model used in this method is

$$Y_i(\tilde{t}) = W_{i1}\gamma_1 + W_{i2}\gamma_2 + \zeta(\tilde{t}) + Y_i(1)\beta(\tilde{t},1) + Y_i(r)\beta(\tilde{t},r) + Z_i(1)\delta(\tilde{t},1) + Z_i(r)\delta(\tilde{t},r) + \epsilon_i(\tilde{t}), \quad (2)$$

where W_{i1} and W_{i2} are scalar covariates, $\zeta(\tilde{t})$ is the intercept, $Y_i(\tilde{t})$ represents the response at a particular time point $\tilde{t}$, $Y_i(1)$ and $Y_i(r)$ are the 1st and r^{th} observations respectively from historical data curves, $Z_i(1)$ and $Z_i(r)$ are the 1st and r^{th} observation of covariates respectively. Model parameters β and δ are the corresponding coefficients for the Y and Z historical data.

The goal of this method is to estimate the model parameters and provide $\hat{Y}_i(\tilde{t})$ as dynamic predictions. The parameters in BENDY can be estimated by the linear model (lm)

method in R. Model (2) is used to predict the scalar response $Y_i(\tilde{t})$ and obtain the corresponding prediction interval. Outliers would be those observations that are outside of the BENDY prediction interval. The BENDY model uses only the first and the last observation of available historic data for the Y and Z stochastic processes, whereas DPFFR uses all the available data for time points $1,2,\ldots,r$ for the prediction. BENDY is used to predict a scalar response, while DPFFR is used to predict functional responses.

3.2 DLM

Dynamic Linear Model (DLM) is similar to BENDY in the sense that DLM also predicts a scalar response instead of a functional response, but DLM uses all the available observations from the 1^{st} and the r^{th} time point for the prediction. The model is given by

$$Y_i(\tilde{t}) = W_{i1}\gamma_1 + W_{i2}\gamma_2 + \zeta(\tilde{t}) + \sum_{j=1}^{r} Y_i(t_j)\,\beta(\tilde{t}, t_j) + \sum_{j=1}^{r} Z_i(t_j)\delta(\tilde{t}, t_j) + \epsilon_i(\tilde{t}). \quad (3)$$

In model (3), $Y_i(\tilde{t})$ represents the scalar response at $\tilde{t}$, W_{i1} and W_{i2} are scalar covariates, $\zeta(\tilde{t})$ is an intercept, and $\epsilon_i(\tilde{t})$ is an error term. $\beta(\tilde{t}, t_j)$ and $\delta(\tilde{t}, t_j)$ are parameters that are estimated at $\tilde{t}$ and t_j. The $Y_i(t)$ observations at time t_j which are more important for predicting response would have larger value of parameter $\beta(\tilde{t}, t_j)$. Similarly, when the covariate data at time point t_j $Z_i(t_j)$ are more important for the prediction then $\delta(\tilde{t}, t_j)$ would become larger.

DLM method uses the same covariates as DPFFR. Similar to DPFFR, DLM includes all the available data from $1,2,\ldots,r$ and scalar covariates for prediction of the response. Instead of integration $\int_T Y_i(t)\beta(\tilde{t}, t)dt$, DLM uses sums over discrete time points t_j. In other words, we can say that DPFFR is the functional version of DLM method. On

the other hand, in DPFFR, $\beta(\tilde{t}, t_j)$ and $\delta(\tilde{t}, t_j)$ are taken to be smooth surfaces, but the DLM model does not have such requirement. The parameters in DLM can be estimated by the linear model (lm) method in R (R Core Team 2017). DLM uses an unpenalized least squares criterion. In R, the 'lm' method is used to fit the DLM model and the corresponding predict function is used to construct the prediction interval. Outliers are detected based on the prediction interval.

3.3 DPFR

Dynamic Penalized Functional Regression (DPFR) is described by model (1). It is similar to DPFFR, and one difference is that $Y_i(\tilde{t})$ is considered to be scalar. We can predict only one point in DPFR by using scalar-on-function regression thus, prediction is done at every point $\tilde{t}$. Therefore, to obtain the entire curve one needs to predict at each point $\tilde{t}$ in turn. This implies that the bivariate parameters $\beta(\tilde{t}, t_j)$ and $\delta(\tilde{t}, t_j)$ would be re-fit for each time point in $\tilde{t}$. We use the refund (Goldsmith et al. 2017) and mgcv (Wood 2017) R packages for implementation of DPFR.

3.4 Dynamic_FLR

Dynamic Functional Linear Regression (Shang, 2015, Section 4.6) is a dynamic prediction method for functional data which considers only available data from the Y process for the prediction. Dynamic_FLR can be used to predict a functional response, and do not use functional or scalar covariates to make predictions. The method relies on the functional principal components decomposition of the Y process. We use R software for the implementation in our dataset by employing the dynamic_FLR function in ftsa

(Hyndman and Shang, 2016) R package. Dynamic_FLR also constructs prediction intervals.

3.5 Dynupdate

Dynamic prediction by penalized least squares (Shang, 2015, Section 4.4) is similar to dynamic_FLR as both methods can be used to predict a functional response, and do not use functional or scalar covariates to make predictions. The model for dynupdate uses eigenfunctions of the Y process and penalized coefficients. We use the dynupdate function ftsa (Hyndman and Shang, 2017) library from R software for the implementation. Bootstrap is used to derive prediction intervals in dynupdate.

3.6 Methods for identification of outliers

In general, an outlier is an observation that deviates considerably from other observations as to arouse suspicion that it was generated by a different mechanism (Hawkins, 1980). Different methods were proposed in recent years for the identification of functional outliers in functional data analysis. Febrero et al., 2008 proposed a nonparametric method for functional outlier detection based on the concept of functional depth. Depth measures centrality of a point providing a way to order points in Euclidean space from the center of data to outward. High depth means centrality. The corresponding curve of functional outlier has significantly lower depth. Therefore, by finding the curve with lower depth one can detect functional outliers. The first step of the proposed method (Febrero et al. 2008) for the identification of functional outliers in a given dataset of functional curves is to obtain the functional depths for each curve. If the depth measure is

below a value D, then assume that curves with depth smaller than D are outliers. Febrero et al., 2007 developed another algorithm based on distance for identification of functional outlier.

The R package fda.usc (Febrero-Bande et al., 2012) has several implementations in order to identify functional outliers. We use two methods for functional outlier detection available in R. The first function is $outliers.depth.trim$ (Febrero et al., 2008) and the second function is $outliers.depth.pond$ (Febrero et al., 2008). Both functions use bootstrap to find a distribution of possible D values followed by choosing the median of the resulting bootstrap distribution.

- The method of outliers.depth.trim is the outlier detection method that corresponds to the approach of Febrero et al., 2008 using the functional depth. A number of B bootstrap samples are constructed. Each bootstrap sample consists of a random selection of curves with replacement. For each bootstrap sample, the value of D for that bootstrap sample is set as the empirical 1% percentile of the distribution of depths calculated for the curves in the bootstrap sample.
- The method of outliers.depth.pond is the outlier detection method that corresponds to the approach of Febrero et al., 2008 using the functional depth for each curve. The bootstrap sample is constructed by taking into account the selection of each curve according to functional depth. The bootstrap selects the curves with probabilities proportional to their depth. A higher chance of selecting a curve with high depth is achieved. The first empirical percentile of the distribution of depths from the bootstrap sample of curves is taken as the value for D corresponding to the bootstrap sample.

4 Research Goal

The goal of this work is to observe the comparison between the dynamic prediction interval and the observed trajectory in order to identify departures from the dynamic prediction. We study the effectiveness of DPFFR method and compare this method with several other methods including BENDY, DLM, DPFR, dynamic_FLR, and dynupdate. For BENDY, DLM, DPFR, DPFFR, dynamic_FLR, and dynupdate we use leave one-curve out cross validation to obtain dynamic predictions and dynamic prediction intervals. We study several applications of the methods for detecting functional outliers in functional data. For numerical studies we simulate data with different choices on the number of subjects (n) and using different lengths (r). We apply these methods to real data on Humidity and Temperature to detect dynamic outliers in environmental settings. For the analysis and implementation of these methods we use R software.

5 Numerical Results

5.1 Simulation design

For simulated functional data, we consider the data setting generated from a dynamic functional regression model (1). The bivariate functional parameter is $\beta(\tilde{t}, t) = \cos(2\tilde{t}\pi/16)\sin(2t\pi/16)$, for $\tilde{t} \in \{r+1, r+2, \ldots, M\}$ and $t \in \{1, 2, \ldots, r\}$. The composition of the functional parameter $\beta(\tilde{t}, t)$ is inspired by the form of a bivariate parameter considered in Ivanescu et al. (2015). The functional intercept is $\zeta(\tilde{t}) = e^{-(\tilde{t}-12.5)^2}$ and the random errors $\epsilon_i(\tilde{t})$ were simulated i.i.d. $N(0, 0.22^2)$. Scalar covariates

were generated as $W_1 = 1\{Unif[0,1] > 0.75\}$ and $W_2 \sim N(0, 0.1^2)$, while their corresponding scalar effects were simulated as $\gamma_1 = 1$ and $\gamma_2 = -0.5$. The presence of two scalar covariates, a continuous and a binary covariate, is aligned with models discussed in Ivanescu et al. (2015). By approximating the integrals via Riemann sums with a dense grid for domain t, n functional responses $Y_i(\tilde{t})$ have been generated from model (1), $1 \leq i \leq n$.

For the functional data $Y_i(t)$ we consider the following mean zero process $Y_i(t) = \sum_{k=1}^{10}\{\rho_{ik} \sin\left(\frac{2k\pi t}{10}\right) + p_{ik} \cos\left(\frac{2k\pi t}{10}\right)\}$, where $\rho_{ik}, p_{ik} \sim N\left(0, \frac{1}{k^2}\right)$ are independent across subjects $i, i = 1, 2, \ldots, n$. For the functional predictor $Z_i(t)$ we considered $Z_i(t) = \sum_{k=1}^{40}\left(\frac{2\sqrt{2}}{k\pi}\right) U_{ik} \sin\left(\frac{k\pi t}{16}\right)$, and where $U_{ik} \sim N(0,1)$.

Here we consider the following choices.

i) Number of subjects $n = 25$ and $n = 50$.

ii) Effects for $\delta(\tilde{t}, t)$: We considered two different settings, setting A and setting B for the functional parameter δ. The two setting corresponds to two different forms of the bivariate model parameter.

$$\text{Setting A. } \delta(\tilde{t}, t) = \sqrt{t} \sin\left(\frac{2\tilde{t}\pi}{16}\right)/4.2$$

$$\text{Setting B. } \delta(\tilde{t}, t) = \sqrt{t\tilde{t}}/4.2$$

iii) Number of points at which we have data for all curves $r = 8 \; and \; r = 11$.

iv) Type of outliers: dynamic global functional outliers where outlier effect is $OU = 1$ and $OU = 2$.

After generating the dynamic prediction $\hat{Y}_i(\tilde{t})$ for DPFFR, we construct approximate 95% prediction intervals for DPFFR predictions as

$$\hat{Y}_i(\tilde{t}) \pm t^*_{M-r}\sqrt{var\{\hat{Y}_i(\tilde{t})\} + var\{\epsilon_i(\tilde{t})\}},$$

where t^*_{M-r} is the t quantile with $(M-r)$ degrees of freedom and corresponding to a confidence level 95%. The term $var\{\hat{Y}_i(\tilde{t})\}$ is the variance of the predicted response at time $\tilde{t}$ and $var\{\epsilon_i(\tilde{t})\}$ is the variance of the error term at time point $\tilde{t}$.

5.2 Metrics

For each method that produces dynamic predictions and prediction intervals we use several metrics for prediction performance and identification of outliers. The metrics we use here include Integrated Mean Prediction Error (IMPE), the detection frequency of dynamic outliers, and the distance of outliers from the prediction interval. IMPE is the mean squared error of prediction over the time points of prediction ($\tilde{t}$). We detect dynamic outliers by investigating if the future curve $Y_i(\tilde{t})$ at point $\tilde{t}$ falls outside the dynamic prediction interval.

IMPE is defined as the sum of squared differences between observed value and the predicted value of responses. IMPE is given by

$$IMPE = \frac{1}{n} * \frac{1}{M-r}\sum_{i=1}^{n}\sum_{j=r+1}^{M}\left(Y_i(\tilde{t}_j) - \hat{Y}_i(\tilde{t}_j)\right)^2,$$

where the time points $t_j \in \{r+1, r+2, \ldots M\}$, M is the total number time points, r is the cutoff point, n is the total number of curves. $Y_i(\tilde{t}_j)$ is the observed response at the j^{th} time interval of subject i in the process, $\hat{Y}_i(\tilde{t}_j)$ represents the predicted value of i^{th} observation at time point $\tilde{t}_j$. IMPE is the average of the sum of the squared differences between the observed and predicted responses. The higher the value of IMPE the higher the error of

the prediction for the future response. This means that the method with small IMPE is preferred.

For detection of outliers we use Mean Identification Frequency (MIF). MIF for curve i is defined as

$$MIF_i = \frac{1}{M-r}\sum_{j=1}^{M-r} 1\{Y_i(\tilde{t}_j) \notin [LB_i(\tilde{t}_j), UB_i(\tilde{t}_j)]\},$$

where $Y_i(\tilde{t}_j)$ is the observed response at the j^{th} time interval of subject i in the data. M is the total number of time points, r is the cutoff point, $M-r$ is the number of time points in the interval. We use the indicator function to obtain a value of 1 if data is outside the prediction interval, and 0 if it is inside. For calculating MIF we take the sum of the indicator functions calculated for each $\tilde{t}$ in the interval, then divide by the total number of points in the interval. Curve i has $MIF = 1$ when the curve $Y_i(\tilde{t}_j)$ is outside the bound for all $\tilde{t}_j$. $LB_i(\tilde{t}_j)$ and $UB_i(\tilde{t}_j)$ are the lower and upper bound of the 95% prediction interval for the i^{th} curve at j^{th} time point respectively. The term $IAI = \frac{1}{n}\sum_{i=1}^{n} MIF_i$ is the integrated actual identification. A value of IAI of 1 implies that identification of dynamic outliers had occurred for all points $\tilde{t}_j$ and for all n curves of simulated outliers.

If the dynamic outlier point is detected, we can define mean distance (MD) of outliers from the interval for the i^{th} curve as

$$MD_i = \frac{1}{M-r}\sum_{j=1}^{M-r} \min\{(Y_i(\tilde{t}_j) - LB_i(\tilde{t}_j))^2, (Y_i(\tilde{t}_j) - UB_i(\tilde{t}_j))^2\}.$$

MD can be described as the average of the minimum of the square of distance of each data point $Y_i(\tilde{t}_j)$ from the corresponding bounds of the prediction interval. The term $IDIST = \frac{1}{n}\sum_{i=1}^{n} MD_i$ is the integrated distance. IDIST is a measure of distance between the outlier and the prediction interval bounds.

The mean width (MW) for the i^{th} curve is given by

$$MW_i = \frac{1}{M-r}\sum_{j=1}^{M-r}\left(UB_i(\tilde{t}_j) - LB_i(\tilde{t}_j)\right),$$

where $\tilde{t}_j \in \{r+1, r+2, \ldots, M\}$, and r is the cutoff point. $UB_i(\tilde{t}_j)$ is the upper bound of the 95% prediction interval for the i^{th} curve at time point $\tilde{t}_j$, $LB_i(\tilde{t}_j)$ is the lower bound of the 95% prediction interval for the i^{th} curve at time point $\tilde{t}_j$. The MW can be described as the average width of each dynamic prediction interval generated by a dynamic prediction method for curve i. The integrated average width (IAW) is the average of the mean width across all curves and is given by $IAW = \frac{1}{n}\sum_{i=1}^{n} MW_i$.

5.3 Simulation of data

We use R software for the simulation of functional data having a total number of time points $M = 16$, sample size for the number of curves $n = 25, 50$, and cutoff points $r = 8, 11$. We use model (1) for the simulation. Two simulated instances of functional datasets have been shown in Figure 1.

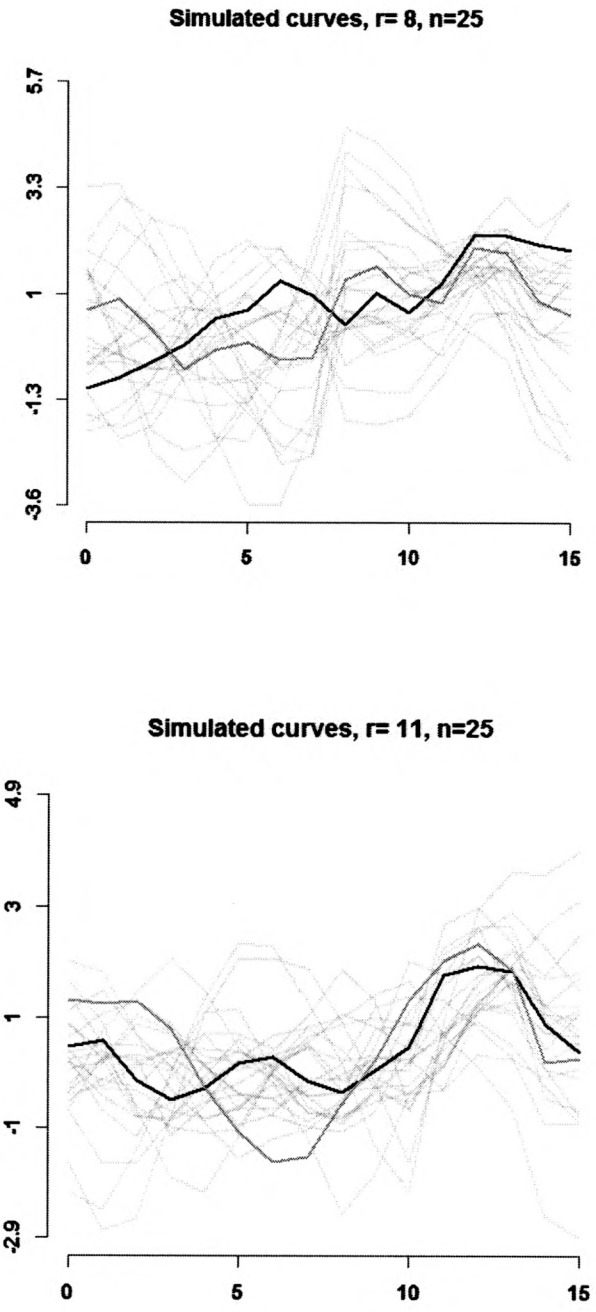


Figure 1: Sample of 25 simulated curves for setting A with cutoff point $r = 8$ (panel in the first row) and $r = 11$ (panel in the second row).

Figure 1 shows sample of 25 simulated curves with different cutoff points r, where the trend of two particular curves have been shown in black and dark gray solid lines. The remainder curves are illustrated in gray.

5.4. Simulation of outliers

After functional data was simulated, the outliers were simulated. For dynamic outliers we considered $OU = 1$ and $OU = 2$. The outliers were simulated as dynamic global outliers. The *i*-th simulated dynamic outlier was a functional outlier that had data values given by $Y_i(\tilde{t}) + OU$ for each $\tilde{t}$. We considered two data generation scenarios scenarios for $Y_i(\tilde{t})$, one without scalar covariate and another one with scalar covariates in the model.

For the construction of prediction intrevals we used the leave one curve-out cross validation method. Therefore, the methods use models fitted on a sample of n-1 curves. The steps for the simulation of dynamic outliers and the method of identification are outlined.

- Step 1. Simulate data $Y_i(\tilde{t})$ according to the DPFFR model.
- Step 2. In Step 2 the simulated dynamic outlier is generated. For curve *i* (one curve) the values for the functional responses were modified as $Y_i(\tilde{t}) + OU$ at each point $\tilde{t}$.
- Step 3. The remaining curves (a sample of *n-1* curves) were used for model fitting.
- Step 4. The methods of prediction and construction of prediction intervals were applied with the historic data for curve *i* and the model fit from Step 3. A prediction interval for curve *i* was obtained.

- Step 5. The simulated outlier was identified as a dynamic outlier at $\tilde{t}$ if at point $\tilde{t}$ it fell outside the dynamic prediction interval.
- Step 6. Steps 2-5 were repeated for each curve *i* in turn.
- Step 7. Metrics were computed by taking into account the sample of *n* simulated outliers.

5.5 Results in simulations

Results are presented for two cases: (i) Model with two functional predictiors and no scalar covariates, and (ii) Model with two functional predictiors and two scalar covariates.

(i). Model with two functional predictiors and no scalar covariates :

We first considered the case with model

$$Y_i(\tilde{t}) = \zeta(\tilde{t}) + \int_T Y_i(t)\beta(\tilde{t},t)dt + \int_T Z_i(t)\delta(\tilde{t},t)dt + \epsilon_i(\tilde{t}),$$

where the model had two functional covariates $Y_i(t)$ and $Z_i(t)$ and no scalar covariates. Specifications for model components were the same as in model (1).

The results obtained by different methods for detection of dynamic outliers with two functional predictors and no scalar covariates with size of outliers $OU = 1$ and $OU = 2$ are given in Table 1 and Table 2. These tables show five metrics IMPE, IAI, IDIST, IAW, and cpu.

Table 1 : Results for dynamic outliers identification. Number are averages across 100 simulations. Scenario: Two functional predictors and no scalar covariates, dynamic global outlier $OU = 1$.

		r = 8					r = 11				
	Method	IMPE	IAI	IDIST	IAW	cpu	IMPE	IAI	IDIST	IAW	cpu
Setting A											
$n = 25$	BENDY	0.15	0.70	0.26	1.58	0.98	0.52	0.33	0.39	2.84	0.59
	DLM	0.16	0.63	0.27	1.76	1.40	0.55	0.24	0.50	3.97	1.03
	DPFR	0.12	0.86	0.29	1.29	10.03	0.26	0.67	0.30	1.72	5.85
	DPFFR	0.09	0.91	0.28	1.16	5.89	0.07	0.92	0.22	1.27	4.45
	dynamic_FLR	0.17	0.47	0.25	2.29	10.85	0.48	0.26	0.43	3.35	9.01
	dynupdate	0.10	0.98	0.59	0.63	13.75	0.27	0.85	0.66	0.92	13.04
$n = 50$	BENDY	0.14	0.76	0.26	1.46	1.96	0.48	0.36	0.35	2.68	1.23
	DLM	0.08	0.94	0.30	1.10	2.79	0.09	0.90	0.28	1.20	2.14
	DPFR	0.91	0.92	0.29	1.16	21.55	0.19	0.77	0.30	1.44	12.73
	DPFFR	0.09	0.93	0.30	1.09	15.88	0.06	0.95	0.23	1.21	11.17
	dynamic_FLR	0.14	0.59	0.23	1.87	316.32	0.39	0.31	0.33	2.93	28.19
	dynupdate	0.12	0.97	0.56	0.70	75.55	0.31	0.81	0.65	1.01	75.77
Setting B											
$n = 25$	BENDY	0.60	0.24	0.44	3.16	0.94	2.42	0.09	1.32	6.33	0.59
	DLM	0.16	0.63	0.27	1.76	1.35	0.55	0.24	0.50	3.97	1.03
	DPFR	1.13	0.48	0.64	2.57	9.20	3.57	0.31	1.22	4.76	5.62
	DPFFR	0.09	0.92	0.28	1.15	6.07	0.07	0.92	0.22	1.27	5.19
	dynamic_FLR	2.34	0.09	1.26	7.15	9.37	7.88	0.07	3.26	12.90	7.64
	dynupdate	1.61	0.31	0.91	3.69	13.13	4.48	0.15	1.69	8.31	13.08
$n = 50$	BENDY	0.55	0.27	0.35	2.95	1.94	2.38	0.10	1.06	6.13	1.23
	DLM	0.08	0.94	0.30	1.10	2.76	0.09	0.90	0.28	1.20	2.15
	DPFR	0.84	0.60	0.68	1.94	20.71	2.61	0.42	1.31	3.38	148.70
	DPFFR	0.09	0.94	0.30	1.09	16.16	0.06	0.95	0.23	1.21	13.06
	dynamic_FLR	2.01	0.09	0.93	6.54	28.87	6.58	0.06	2.56	11.76	23.89
	dynupdate	1.71	0.25	0.91	4.12	75.39	5.82	0.11	1.83	8.87	75.35

Results : Comparisons have been made from the values obtained in several simulated scenarios. We compare changes while

i) cut offf point r was changed from 8 to 11,

ii) number of curve n was changed from 25 to 50,

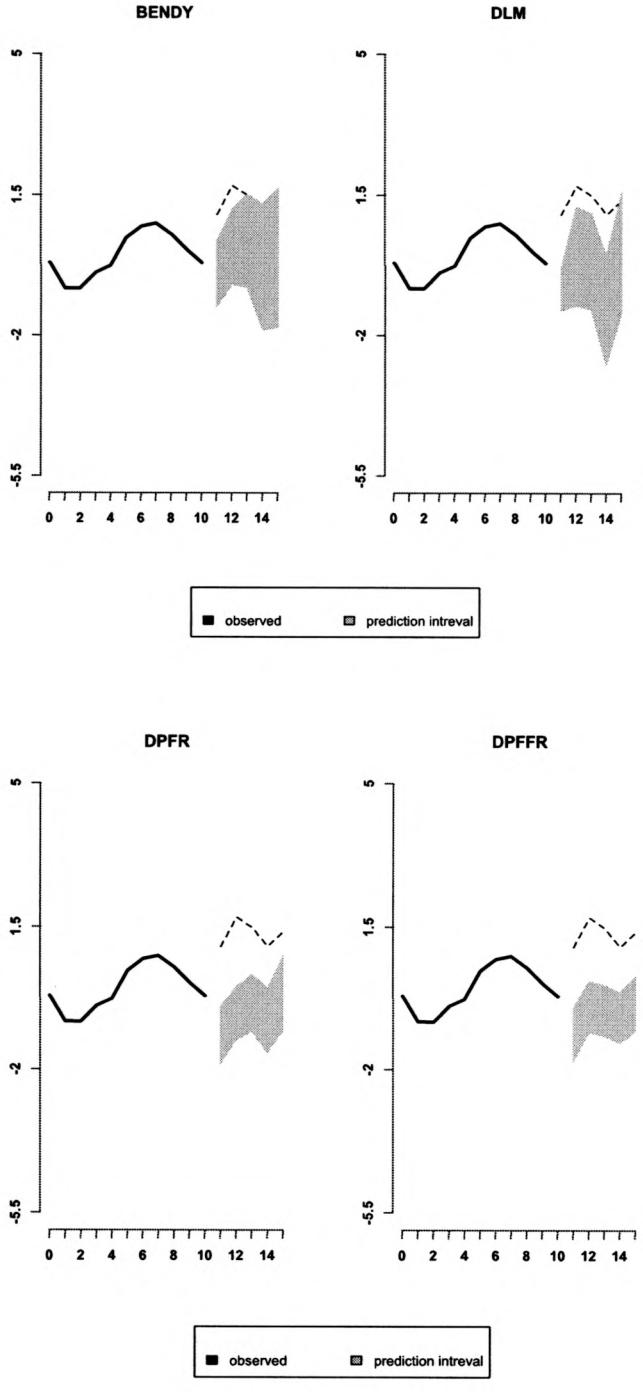
iii) data settings A and B were considered.

We observed that the identification rate became larger or was maintained high for DPFFR when cutoff point increased from 8 to 11. As cutoff point (r) increased, IMPE for DPFFR decreased. The average width increased for all methods as cutoff point increased. For DPFFR the distance from actual outliers became smaller as length (r) of historic data increased. This may be due to the fact that the width of the prediction intervals increased as r increased. There were slight changes in processing time (cpu) when r increases from 8 to 11 where for $r = 11$ the cpu was overall smaller. As a higher number of curves was acquired the rate of identification became higher for DPFFR.

When the number of curves increased, the identification rate of outliers was higher for most of the methods except dynupdate for both setting A and B. IMPE decreased when number of curves increased except dynupdate. When we add more curves, we can get more precise prediction estimates and more easily identify how observed values compare to predicted values. Distance from actual outliers increased as number of curves increased for each method. The average width became smaller as sample size n increased. For DPFFR when the number of curves increased the average width decreased. With higher number of curves, the dynamic outlier identification encumbered increased computation time. Some patterns of changes is obvious for data setting B.

We observed that identification rate of outliers decreased from data setting A to setting B, except DPFFR, where a high rate of identification was maintained. IMPE increased as the data setting changed from A to setting B. There were small differences in distance from actual outliers for setting A and setting B. Data setting B takes similar computation time for the detection of dynamic outliers.

Figures 2 shows results for the identification of dynamic outliers by six methods for setting A when the model used for data generation has two functional predictors and no scalar covariates.



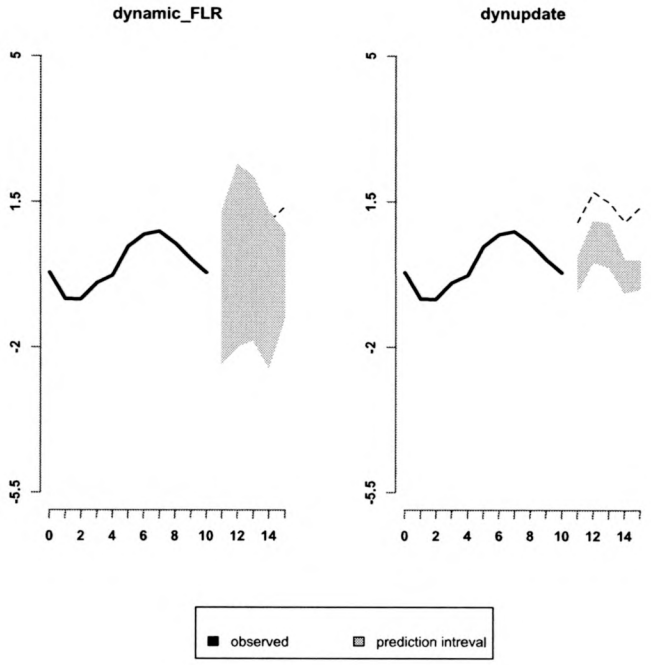


Figure 2: Dynamic outlier identification with setting $A, OU = 2, r = 11$, and $n = 25$. The model has two functional predictors and no scalar covariates.

Figure 2 shows the identification of outliers by six methods with the model having two functional predictors and no scalar covariates, when global outliers $OU = 2$, number of curves $n = 25$, cutoff point $r = 11$, and setting A were considered. The grey regions are the prediction intervals and dashed lines are the simulated dynamic outliers. These graphs show that DPFFR identified dynamic outliers more precisely or as well as other methods.

Table 2 shows the results for global outliers with a size $OU = 2$. We also compare the results between Table 1 and Table 2 to have more insight about the methods for detection of dynamic outliers.

Table 2 : Results for dynamic outliers identification. Number are averages across 100 simulations. Scenario: Two functional predictors and no scalar covariates, dynamic global outlier $OU = 2$.

		r = 8					r = 11				
	Method	IMPE	IAI	IDIST	IAW	cpu	IMPE	IAI	IDIST	IAW	cpu
Setting A											
$n = 25$	BENDY	0.15	1.00	1.64	1.58	0.98	0.52	0.78	1.12	2.84	0.63
	DLM	0.16	0.97	1.53	1.76	1.41	0.55	0.60	1.14	3.97	1.11
	DPFR	0.19	0.99	1.96	1.36	5.93	0.19	0.99	1.82	1.47	3.15
	DPFFR	0.09	1.00	2.12	1.16	8.21	0.07	1.00	1.93	1.27	6.37
	dynamic_FLR	0.17	0.91	1.24	2.28	12.84	0.48	0.68	1.07	3.35	11.38
	dynupdate	0.10	1.00	2.96	0.63	17.21	0.27	0.99	2.68	0.93	17.67
$n = 50$	BENDY	0.14	0.99	1.76	1.46	2.04	0.41	0.87	1.07	2.55	1.20
	DLM	0.08	1.00	2.14	1.14	3.00	0.09	1.00	2.04	1.21	2.25
	DPFR	0.14	1.00	2.11	1.19	13.33	0.16	0.99	1.98	1.31	6.30
	DPFFR	0.09	1.00	2.21	1.09	28.33	0.07	1.00	2.01	1.21	19.73
	dynamic_FLR	0.14	0.99	1.50	1.73	41.53	0.37	0.88	1.11	2.47	37.13
	dynupdate	0.12	1.00	2.79	0.69	102.57	0.29	0.99	2.67	0.98	100.66
Setting B											
$n = 25$	BENDY	0.60	0.70	1.01	3.16	1.19	2.42	0.25	1.85	6.33	0.70
	DLM	0.16	0.97	1.53	1.76	1.72	0.55	0.60	1.14	3.97	1.22
	DPFR	0.37	0.96	1.96	1.57	7.07	0.59	0.88	1.96	1.92	3.65
	DPFFR	0.92	1.00	2.13	1.15	9.89	0.67	1.00	1.93	1.27	8.29
	dynamic_FLR	2.34	0.22	1.97	7.16	12.62	7.88	0.11	4.68	12.91	10.51
	dynupdate	1.61	0.58	1.85	3.69	20.57	5.46	0.24	2.81	8.30	19.48
$n = 50$	BENDY	0.55	0.76	1.01	2.96	1.78	2.38	0.25	1.59	6.13	1.12
	DLM	0.07	1.00	2.18	1.10	2.49	0.09	1.00	2.05	1.20	1.93
	DPFR	0.25	0.99	2.09	1.31	11.14	0.49	0.94	2.10	1.58	5.87
	DPFFR	0.92	1.00	2.21	1.08	14.70	0.60	1.00	2.01	1.21	11.70
	dynamic_FLR	2.01	0.25	1.46	6.53	26.62	6.58	0.11	3.23	11.76	22.01
	dynupdate	1.71	0.51	1.70	4.12	70.85	5.82	0.19	2.68	8.87	70.65

We compared results while changing r from 8 to 11, changing number of curves from 25 to 50, and changing data setting A to B.

With size of outliers $OU = 2$ we observed that identification rate became smaller as the length of the historic data increased from 8 to 11. Only DPFFR maintains maximum rate of identification (100%) when size of the global outliers is $OU = 2$. As length (r)

increased, average width increased while identification rate decreased or stayed about the same. The distance from actual outliers decreased as length (r) of historic data increased and this was due to the fact that the width of the prediction intervals increased. There is only a slight difference in processing time (cpu) when r increases from 8 to 11. Similarly, with higher number of curves rate of identification became slightly smaller or stayed about the same. IMPE increased as cutoff point (r) increased for some methods, while for DPFR and DPFFR the prediction error rate decreased or was maintained small.

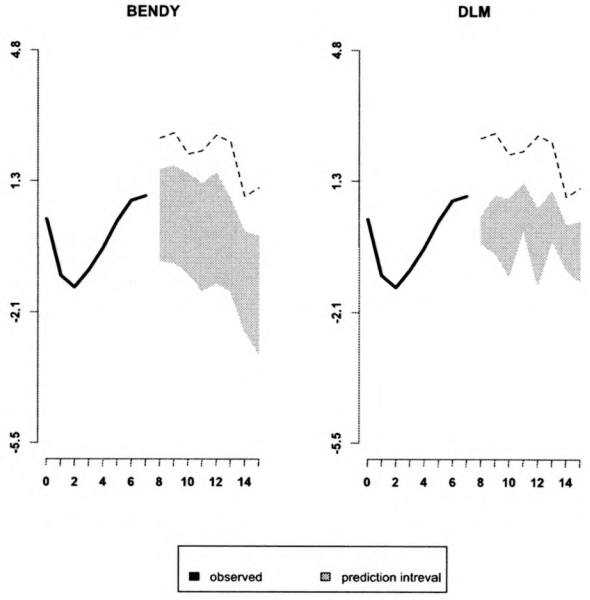
When number of curves increased, identification of outliers was higher for most of the methods. It may be because with a large sample size n the different methods detect outliers more accurately. IMPE decreased when number of curves increased. Distance from actual outliers slightly increased in each method as number of curves increased. On the other hand, the average width decreased as the number of sample increased. For DPFFR when the number of curves increased the average width decreased from 1.16 to 1.09 with cutoff point 8. With higher number of curves, the dynamic outlier identification had increased computation time.

We observed that identification rate of outliers decreased from data setting A to setting B, except DPFFR, where DPFFR maintains maximum rate of identification when there was a large dynamic outlier with $OU = 2$; see IAI results of Table 2. IMPE increased when data setting changed from A to setting B. There were small differences in distance from actual outliers for setting A and setting B. Data setting B takes a little more time for the detection of dynamic outliers.

Comparing Table 1 and Table 2 we observed that the rate of identification and distance from dynamic outliers change when the size of outliers changes from $OU = 1$ to

$OU = 2$. The rate of identification increased when the size of outliers increased. The distance from outliers are greater when the size of outliers is higher, such as when $OU = 2$. From simulated results, we may conclude that detection of outliers change with different $r, n,$ and magnitude of outliers.

Figures 3 shows results for the identification of dynamic outliers by six methods for setting B when the model used for data generation has two functional predictors and no scalar covariates.



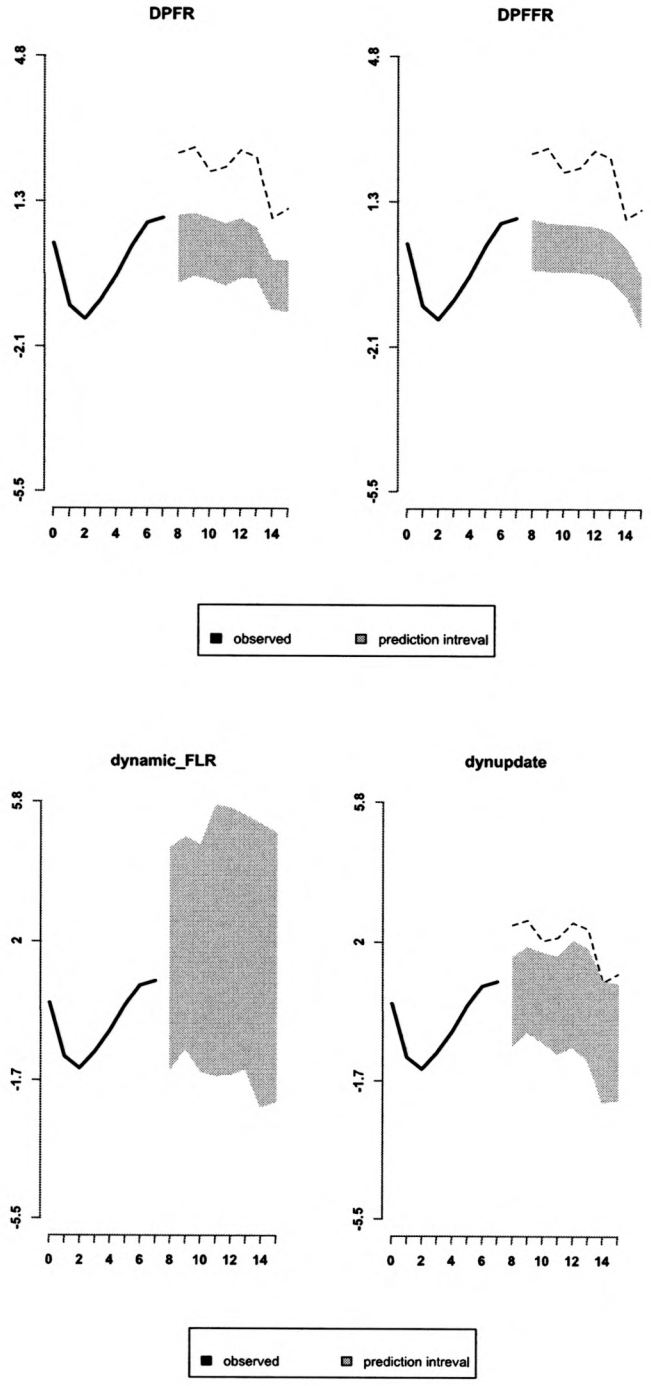


Figure 3: Dynamic outlier identification with setting $B, OU = 2, r = 8$ and $n = 25$. The model has two functional predictors and no scalar covariates.

Figure 3 displays the identification of dynamic outliers by different methods with the model having two functional predictors and no scalar covariates, when global outliers OU=2, number of curves $n = 25$, cutoff point $r = 8$, and setting B were considered. These graphs show that DPFFR, DLM, and DPFR identified dynamic outliers. Since dynamic_FLR and dynupdate have relatively wider prediction interval, these methods have smaller identification rates for setting B.

(ii) Identification of dynamic outliers with two functional predictors and two scalar covariates :

The results obtained by different methods for detection of dynamic outliers with two functional predictors and two scalar covariates in the model with size of outliers $OU = 1$ and $OU = 2$ are given in Table 3 and Table 4 respectively. These tables show five metrics IMPE, IAI, IDIST, IAW, and cpu.

Table 3 : Results for dynamic outliers identification. Numbers are averages across 100 simulations. Scenario: Two functional predictors and two scalar covariates, dynamic global outlier $OU = 1$.

		r = 8					r = 11				
n	Method	IMPE	IAI	IDIST	IAW	cpu	IMPE	IAI	IDIST	IAW	cpu
Setting A											
$n = 25$	BENDY	0.17	0.65	0.26	1.69	1.10	0.57	0.31	0.44	3.00	0.70
	DLM	0.22	0.50	0.29	2.14	1.50	3.93	0.08	2.03	26.66	1.09
	DPFR	0.23	0.76	0.40	1.40	5.21	0.20	0.74	0.31	1.53	2.96
	DPFFR	0.09	0.91	0.28	1.17	5.76	0.07	0.91	0.21	1.30	4.42
	dynamic_FLR	0.17	0.47	0.25	2.28	10.27	0.48	0.26	0.43	3.36	8.76
	dynupdate	0.10	0.98	0.59	0.63	13.15	0.27	0.85	0.66	0.92	13.11
$n = 50$	BENDY	0.14	0.75	0.26	1.50	2.31	0.51	0.34	0.35	2.75	1.42
	DLM	0.08	0.93	0.29	1.14	3.12	0.10	0.87	0.27	1.26	2.27
	DPFR	0.17	0.87	0.37	1.21	11.66	0.13	0.83	0.29	1.33	6.83
	DPFFR	0.09	0.93	0.30	1.10	16.32	0.06	0.95	0.22	1.22	11.48

	dynamic_FLR	0.14	0.59	0.23	1.88	32.49	0.39	0.32	0.33	2.92	28.28
	dynupdate	0.12	0.97	0.56	0.70	77.01	0.31	0.82	0.65	1.01	75.44
Setting B											
$n = 25$	BENDY	0.68	0.22	0.49	3.38	1.11	2.66	0.09	1.59	6.68	0.75
	DLM	0.22	0.50	0.29	2.14	1.51	3.93	0.08	2.03	26.66	1.18
	DPFR	0.34	0.72	0.48	1.51	5.44	0.36	0.64	0.43	1.76	3.22
	DPFFR	0.09	0.91	0.28	1.16	6.13	0.07	0.91	0.21	1.29	5.71
	dynamic_FLR	2.34	0.09	1.21	7.16	9.06	7.88	0.07	3.34	12.97	8.22
	dynupdate	1.61	0.31	0.90	3.69	13.13	5.47	0.15	1.67	8.30	30.87
$n = 50$	BENDY	0.58	0.26	0.37	3.03	2.26	2.50	0.10	1.07	6.29	1.45
	DLM	0.08	0.93	0.29	1.14	3.09	0.10	0.87	0.27	1.26	2.32
	DPFR	0.23	0.80	0.43	1.28	11.99	0.26	0.74	0.40	1.46	6.81
	DPFFR	0.09	0.93	0.30	1.09	16.86	0.06	0.95	0.23	1.21	13.81
	dynamic_FLR	2.01	0.09	0.97	6.55	28.99	6.58	0.06	2.59	11.75	24.89
	dynupdate	1.70	0.25	0.88	4.11	75.44	5.83	0.11	1.66	8.85	284.00

Table 4 : Results for dynamic outliers identification. Numbers are averages across 100 simulations. Scenario: Two functional predictors and two scalar covariates, dynamic global outlier $OU = 2$.

		r = 8					r = 11				
	Method	IMPE	IAI	IDIST	IAW	cpu	IMPE	IAI	DIST	IAW	cpu
Setting A											
$n = 25$	BENDY	0.17	0.99	1.55	1.69	1.52	0.57	0.75	1.13	3.00	0.79
	DLM	0.22	0.93	1.32	2.14	1.52	3.93	0.14	3.31	26.66	1.25
	DPFR	0.23	0.99	1.97	1.40	5.25	0.20	0.98	1.77	1.53	3.29
	DPFFR	0.09	1.00	2.09	1.17	5.77	0.07	1.00	1.90	1.30	5.00
	dynamic_FLR	0.17	0.91	1.24	2.28	10.14	0.48	0.68	1.06	3.36	9.70
	dynupdate	0.10	1.00	2.96	0.63	13.17	0.27	1.00	2.68	0.92	14.57
$n = 50$	BENDY	0.14	1.00	1.72	1.50	2.28	0.51	0.80	1.13	2.75	1.40
	DLM	0.08	1.00	2.13	1.14	3.11	0.09	1.00	1.99	1.26	2.25
	DPFR	0.17	1.00	2.12	1.21	11.56	0.13	1.00	1.93	1.33	6.67
	DPFFR	0.09	1.00	2.19	1.09	16.09	0.06	1.00	1.99	1.22	11.33
	dynamic_FLR	0.14	1.00	1.43	1.88	31.95	0.39	0.77	1.05	2.92	27.65
	dynupdate	0.12	1.00	2.85	0.69	76.22	0.31	0.99	2.57	1.01	74.47
Setting B											
$n = 25$	BENDY	0.68	0.65	1.03	3.38	1.12	2.66	0.23	2.12	6.68	0.71
	DLM	0.22	0.93	1.32	2.14	1.53	3.13	0.14	3.31	26.66	1.13

	DPFR	0.34	0.96	1.99	1.51	5.98	0.36	0.94	1.78	1.76	3.06
	DPFFR	0.09	1.00	2.10	1.16	6.16	0.07	1.00	1.91	1.29	5.38
	dynamic_FLR	2.34	0.21	1.93	7.16	9.00	7.88	0.11	4.49	12.97	7.68
	dynupdate	1.61	0.58	1.87	3.69	13.20	5.47	0.24	2.79	8.30	13.28
$n = 50$	BENDY	0.58	0.74	1.01	3.03	2.24	2.80	0.24	1.64	6.29	1.40
	DLM	0.08	1.00	2.13	1.14	3.07	0.09	1.00	1.99	1.26	2.26
	DPFR	0.23	0.99	2.20	1.28	12.04	0.26	0.98	1.92	1.46	6.22
	DPFFR	0.09	1.00	1.50	1.09	16.66	0.06	1.00	2.00	1.21	13.36
	dynamic_FLR	2.01	0.24	1.70	6.55	28.50	6.58	0.01	3.33	1.75	23.65
	dynupdate	1.09	0.51	1.70	4.11	74.76	5.83	0.19	2.60	8.85	74.49

With two functional predictors and two scalar covariates when the size of outliers $OU = 1$ and $OU = 2$ (Table 3 and Table 4), we observed that identification rate became smaller or stayed about the same when cutoff point increased from 8 to 11. Only DPFFR maintains the same very high rate of identification (100%). As length (r) increased, IMPE and average width of the dynamic prediction interval also increased. As the width of the prediction interval increased as r increased, the distance from actual outliers decreased as length (r) of historic data increased. There is little difference in computation time when r increased from 8 to 11.

Comparing Table 2 and Table 4 we observed that when two scalar covariates are added there is overall little change in rate of identification, IMPE, and distance. DPFFR had high rate of identification. Average prediction interval widths are slightly higher when models include two functional predictors and two scalar covariates compared to the models with two functional predictors and no scalar covariates.

We also implement functional data analysis methods for functional outlier identification. The R library fda.usc (Febero et al., 2012) includes routines outliers.depth.trim and outliers.depth.pond for identifying outlying functional samples in a

functional dataset using methods from Febrero et al. (2007), Febrero et al. (2008), and Febero et al. (2012). For a given curve i simulated by model (1) a corresponding simulated dynamic outlier is constructed using magnitude OU. The Febrero et al. (2008) method is used to identify if the simulated outlying curve is identified as an outlier in the region $\tilde{t}$. The identification rates with size of outliers $OU = 1$ and $OU = 2$ for cutoff points $r = 8$ and $r = 11$ with two functional predictors and no scalar covariate have been shown in Table 5 .

Table 5: Identification rate by different functional data analysis R functions with $OU = 1$ and $OU = 2$ for cutoff points $r = 8$ and $r = 11$. Scenario: data simulated with two functional predictors and no scalar covariate.

	$OU = 1$		$OU = 2$	
	Identification Rate at $r = 8$	Identification Rate at $r = 11$	Identification Rate at $r = 8$	Identification Rate at $r = 11$
Setting A				
$n = 25$				
outliers.depth.trim	0.03	0.03	0.10	0.06
outliers.depth.pond	0.03	0.03	0.10	0.05
$n = 50$				
outliers.depth.trim	0.01	0.01	0.05	0.03
outliers.depth.pond	0.01	0.01	0.05	0.02
Setting B				
$n = 25$				
outliers.depth.trim	0.05	0.02	0.11	0.04
outliers.depth.pond	0.04	0.02	0.10	0.03
$n = 50$				
outliers.depth.trim	0.03	0.02	0.07	0.03
outliers.depth.pond	0.02	0.02	0.06	0.02

Table 6 shows identification rates with size of outliers $OU = 1$ and $OU = 2$ for cutoff points $r = 8$ and $r = 11$ with two functional predictors and two scalar covariates.

Table 6: Identification rate by different functional data analysis R functions with $OU = 1$ and $OU = 2$ for cutoff points $r = 8$ and $r = 11$, Scenario: data simulated with two predictors and two scalar covariates.

	$OU = 1$		$OU = 2$	
	Identification Rate at $r = 8$	Identification Rate at $r = 11$	Identification Rate at $r = 8$	Identification Rate at $r = 11$
Setting A				
$n = 25$				
outliers.depth.trim	0.03	0.03	0.09	0.05
outliers.depth.pond	0.03	0.03	0.10	0.06
$n = 50$				
outliers.depth.trim	0.01	0.01	0.05	0.03
outliers.depth.pond	0.01	0.01	0.05	0.02
Setting B				
$n = 25$				
outliers.depth.trim	0.04	0.03	0.10	0.04
outliers.depth.pond	0.04	0.02	0.10	0.04
$n = 50$				
outliers.depth.trim	0.03	0.02	0.07	0.03
outliers.depth.pond	0.02	0.02	0.06	0.02

Table 5 and Table 6 show the rate of detection of outliers obtained from functions outliers.depth.trim and outliers.depth.pond (Febrero et al., 2008). These two methods seem to have a little chance to identify dynamic functional outliers. Comparing OU=1 and OU=2 , the rate of detection increases when the size of outliers OU increases. In Table 5 as r increased the rate of identification for Febrero et al. (2008) methods decreased. When two scalar covariates were added to the data generation model the rate of

identification of outliers was similar as for the case of two functional predictors and no scalar covariates for the data generation mechanism.

6 Data Analysis

6.1 Application to real data study

In our study, we have data available for Humidity (in percent) and Temperature (in degrees Celsius) which are part of a dataset from the UCI Machine Learning Repository. In this study, the data set relied on historical data from years 2011 and 2012 in Washington DC.

Humidity and temperature data were collected for a sample of nonconsecutive $n = 75$ days, hourly for each day. We can store a functional dataset, such as the humidity dataset as a 75×24 matrix with 75 rows corresponding to the number of days available and 24 columns corresponding to the number of observations taken hourly within the day. In graphical representation, we have 75 curves where each curve consists of 24 equally spaced measurements for the 24 time points. For example, in this study we denote $Y_i(t)$ as the observed data on Humidity on the equally spaced grid $\{1,2, \dots, r\}$. Moreover, $Z_i(t)$ are data on Temperature which we consider as a functional predictor observed at the same grid of time points. Modeling is done by taking $Y_i(\tilde{t})$ to represent the functional responses of Humidity for the remainder of the day starting with time point $r + 1$ until the end of day at time 24, with $\tilde{t} \epsilon \{r + 1, \dots 24\}$. Figure 4 gives a visual description of the actual dataset of Humidity that is the focus of the dynamic outlier identification analysis.

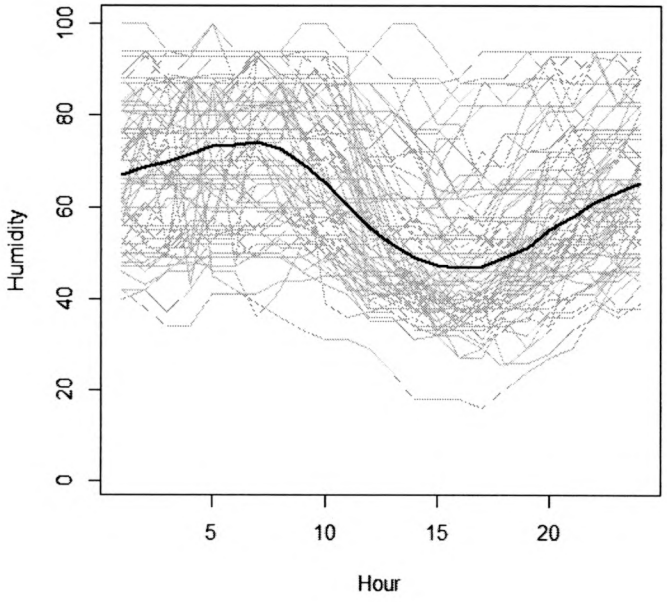


Figure 4: Observed data of Humidity have been shown in grey lines. The bold black line is the functional mean curve.

Figure 4 shows the actual data of Humidity for some nonconsecutive 75 days collected for each of the 24 hours each day. The graph illustrates 75 curves that represent the humidity for 75 days where daily humidity was collected hourly at equally spaced time points $\{1, 2, ..., 24\}$. The black line has been obtained by considering the mean at each time point from all 75 curves and also applying a smoothing approach to obtain a smooth functional mean (Ramsay and Silverman 2005, Ramsay et al., 2016). The minimum observed humidity for the dataset was 20% and the maximum humidity was 100%. The data pertains to humidity and temperature observed in Washington D.C. The mean humidity is higher in the beginning and towards the end of the day and lower in the mid of the day.

6.2 Functional principal components

Functional principal component analysis (fPCA) has vital role in the development of FDA. The method of fPCA is the first step to represent the functional data in a lower dimensional space and to capture the main sources of variability of the data by means of a small number of components (Ramsay and Silverman 2005). Therefore, it is important to estimate the principal components that contain and explain most of the variability in a given functional data sample. The functional data $Y_i(t)$ can be decomposed as

$$Y_i(t) = \sum_{k=1}^{\infty} f_{ik}\xi_k(t)$$

where f_{ik} are pairwise uncorrelated random variables, and the functions ξ_k are pairwise orthogonal in τ with $t \in \tau$.

The covariance function of the data is the surface is defined as $Cov\big(Y_i(t), Y_i(s)\big) = V(t,s)$. It is assumed that there is an orthogonal expansion of V in terms of *eigenfunctions* ξ_k and nondecreasing *eigenvalues* d_k.

$$V(t,s) = \sum_{k=1}^{K} d_k\xi_k(t)\xi_k(s)$$

where $\xi_k(t)$ are the functional principal components, or harmonics. This step enables the calculation of $var(\epsilon(t))$ for the data application where the estimated $\hat{Y}(t)$ and observed $Y(t)$ are used in the calculation of the variance.

The first functional principal component (fPC) corresponds to the most important mode of variation, and the second fPC which is orthogonal to the first one corresponds to the second most important mode of variation. These principal components correspond to the eigen functions of the empirical covariance function. We obtained functional principal components from the analysis for Humidity data by using the fpca.sc (Di et al. 2009;

Goldsmith et al. 2013) function in the refund (Goldsmith et al., 2016) R package. This fPCA method uses smoothing splines for the estimation of the covariance.

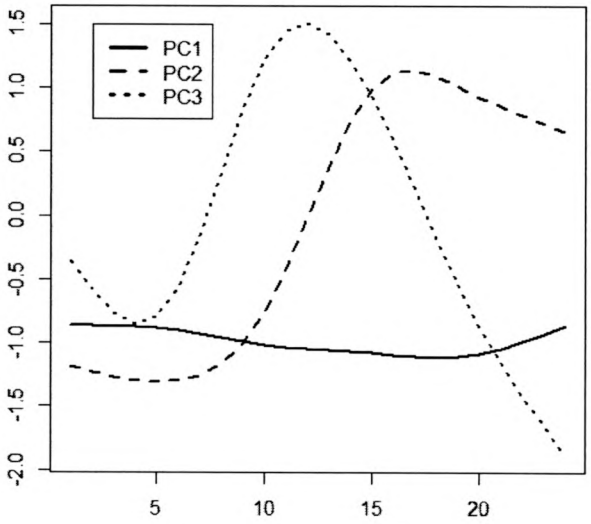


Figure 5: Graph displays the first three estimated functional principal components from a functional principal components analysis (fPCA) for the humidity data.

The black solid line (fPC1) is the functional shape that suggests some vertical main shift of the curves for humidity. Dashed line corresponds to the estimated fPC2 which seems to display a contrast between the first half of the day and the second half of the day. The dotted line is fPC3 which contrasts in the middle with the rest. The percent of explained variability were: 97.3% (fPC1), 1.3% (fPC2), and 0.4% (fPC3).

6.3 Results in application to data analysis

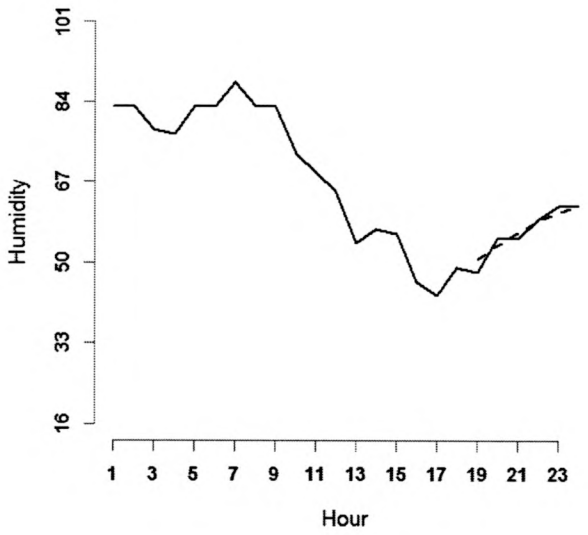
We use DPFFR, dynamic_FLR (Shang, 2015; Hyndman and Shang 2016), and dynupdate (Shang, 2015; Hyndman and Shang 2016) to compare the performance of these

methods when predicting the future trajectory of humidity, followed by identification of humidity data points that are dynamic outliers. For this purpose, we took different time points (r) for the dynamic prediction. We use R software for the analysis. Libraries fda (Ramsay et al., 2016), ftsa (Hyndman and Shang, 2016), and refund (Goldsmith et al., 2016) are used for this analysis. The model we used for DPFFR is

$$Y_i(\tilde{t}) = \zeta(\tilde{t}) + \int_T Y_i(t)\beta(\tilde{t},t)dt + \int_T Z_i(t)\delta(\tilde{t},t)dt + \epsilon_i(\tilde{t})$$

where, $Y_i(\tilde{t})$ and $Y_i(t)$ were the humidity for the second and first half of the day respectively where measurements are taken every hour. $Z_i(t)$ was the temperature for the first half of the day measured hourly. $\zeta(\tilde{t})$ is the functional intercept at time $\tilde{t}$, $\beta(\tilde{t},t)$ and $\delta(\tilde{t},t)$ are bivariate model parameters. The term $\epsilon_i(\tilde{t})$ is the error at time $\tilde{t}$.

Figure 6 depicts the dynamic prediction for humidity by DPFFR method for two different days.



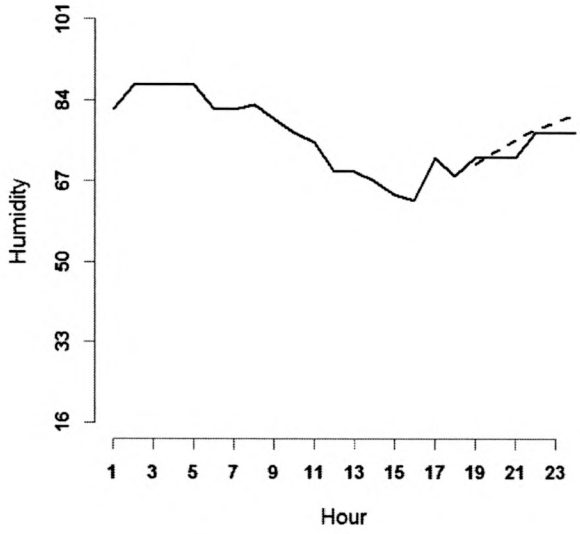


Figure 6: Dynamic prediction by DPFFR method for humidity *with* $r = 18$ for two days. The black line illustrates the actual data for humidity. The dashed black line shows the DPFFR dynamic prediction for the humidity when $r = 18$.

Comparing both panels in Figure 6 it is suggested that DPFFR seems to predict the future observation correctly in both days because the dashed line is closer to the observed data.

Table 7 collects the integrated mean prediction error (IMPE) results for predicting Humidity using DPFFR, dynamic_FLR, and dynupdate.

Table 7: Integrated mean prediction error (IMPE) results for dynamic prediction methods, with $r = 10,15,18$.

Method	IMPE of Humidity
DPFFR	146.60, $r = 10$
	108.82, $r = 15$
	72.79, $r = 18$
Dynamic_FLR	162.68, $r = 10$
	116.56, $r = 15$
	78.15, $r = 18$

dynupdate	$299.47, r = 10$ $218.17, r = 15$ $189.79, r = 18$

IMPE results from Table 7 indicate that DPFFR generates more accurate dynamic predictions than dynamic_FLR and dynupdate. As we use more observed data it predicts future values with less error.

In our further analysis, we report the summary of curve-specific results for Root Mean Square Error (RMSE), identification rate of outliers, distance, width, and see how three different methods can identify them.

For each curve *i,* let Root Mean Square Error (RMSE) be defined as

$$RMSE_i = \frac{1}{M-r}\sum_{j=1}^{M-r}(Y_i(\tilde{t}_j) - \hat{Y}_i(\tilde{t}_j))^{1/2}$$

where, $Y_i(\tilde{t}_j)$ is the actual response at the j^{th} time interval of curve i in the process. $\hat{Y}_i(\tilde{t}_j)$ is the prediction of $Y_i(t_j)$, and $M - r$ is the number of points in the interval $\tilde{t}$ where predictions are obtained. Figure 9 displays the mean squared prediction error calculated as an average across curve at each point $r + 1, r + 2, \ldots M$. We predicted till hour $M = 24$ on the basis of available data up to $r = 19$ and $r = 20$. For each curve i at each point $\tilde{t}$ we also calculate the identification rate, the average distance from the prediction interval, and the average width of dynamic prediction interval.

Table 8 shows the five-number summary of the root mean squared error (RMSE), identification rate, distance, and width obtained by different methods considering $r = 19$ and $r = 20$ for predicting humidity.

Table 8: Results for RMSE, identification rate, distance, and width for different methods with $r = 19$ and $r = 20$ for the humidity data.

	Method	$r = 19$						$r = 20$					
		Min	Q_1	Median	Mean	Q_3	Max	Min	Q_1	Median	Mean	Q_3	Max
RMSE	DPFFR	0.93	3.84	5.14	6.28	7.27	26.53	1.51	3.62	4.91	6.08	7.34	18.10
	DFLR	1.58	4.39	6.27	6.95	8.60	20.31	0.89	4.14	6.56	6.81	8.36	19.57
	dynupdate	2.03	3.99	5.73	6.84	8.35	19.29	1.08	4.11	5.90	6.41	7.90	23.33
Identi-fication Rate	DPFFR	0.00	0.00	0.00	0.11	0.27	1.00	0.00	0.00	0.00	0.08	0.15	0.75
	DFLR	0.00	0.00	0.00	0.05	0.13	0.60	0.00	0.00	0.00	0.03	0.10	0.75
	dynupdate	0.00	0.20	0.60	0.51	0.80	1.00	0.00	0.38	0.75	0.59	0.75	1.00
Distance	DPFFR	0.01	5.11	12.35	28.18	38.31	183.18	0.51	7.21	13.21	22.75	37.11	74.05
	DFLR	0.01	0.87	3.14	14.99	7.72	86.49	0.01	1.53	12.58	20.24	29.77	62.94
	dynupdate	0.05	3.01	9.55	32.60	43.12	214.11	0.05	3.76	10.82	29.06	26.71	380.58
Width	DPFFR	21.45	21.77	22.14	22.49	22.86	26.39	23.27	23.61	24.21	24.50	25.04	30.10
	DFLR	30.55	33.37	34.97	35.12	36.57	42.21	28.37	32.68	35.28	35.41	37.76	43.16
	dynupdate	9.23	9.45	9.58	9.57	9.66	10.01	8.16	8.33	8.41	8.42	8.50	8.78

From the above table, we have seen that with cutoff point $r = 19$ DPFFR has the lowest minimum, Q1, median, mean, Q3 value of RMSE. On the other hand, with $r = 20$ DPFFR has the lowest value of RMSE for Q1, median, mean, Q3, and maximum. Figure 7 gives further information on how RMSE can be used for comparing DPFFR, dynamic_FLR (DFLR) and dynupdate in this application.

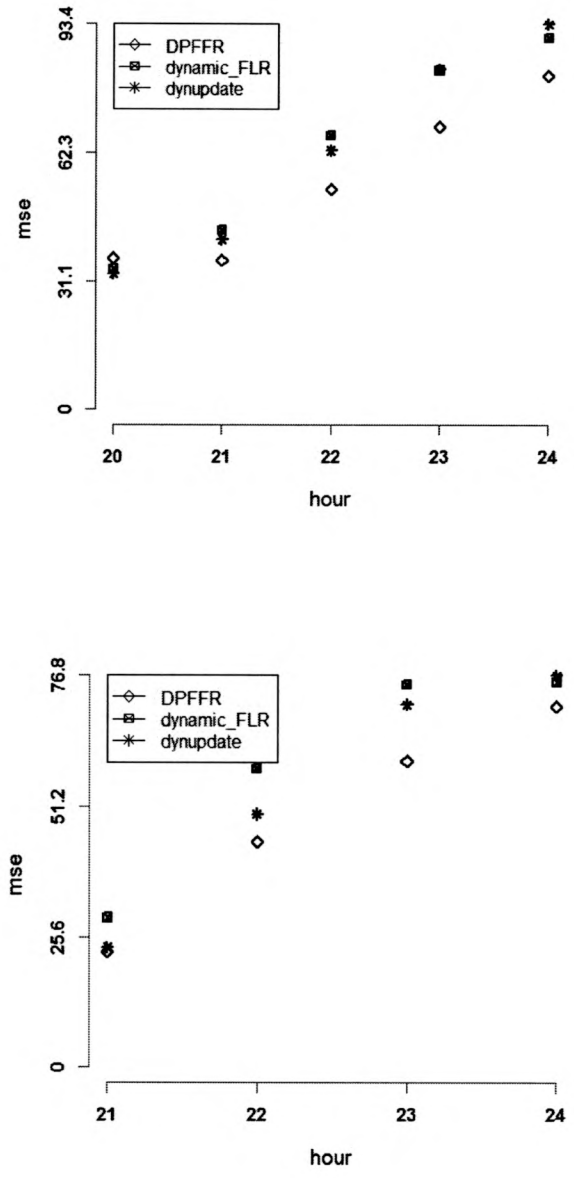


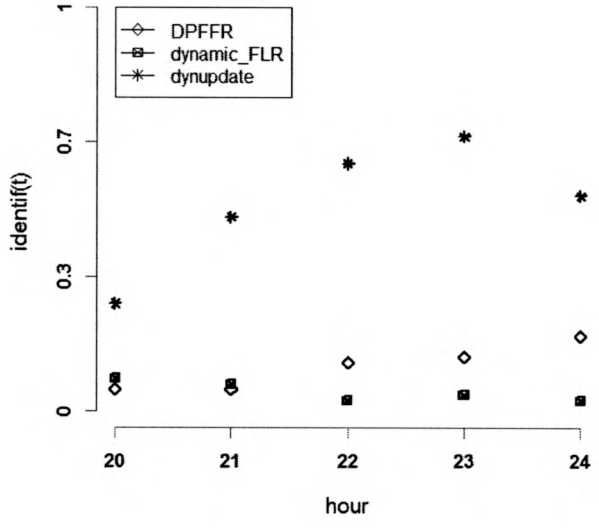
Figure 7: The panel in the first row shows the RMSE by hour when cutoff point is $r = 19$. The panel in the second row shows the RMSE by hour when the cutoff point is $r = 20$.

The above three curves in Figure 7 show that with $r = 19$, at hour 20 RMSE by DPFFR was similar to dynamic_FLR and dynupdate, but time point 21 onwards DPFFR

gives lowest RMSE. We have seen that dynamic_FLR and dynupdate are close to each other. When $r = 20$, RMSE by DPFFR was lower than dynamic_FLR and dynupdate, indicating that DPFFR produced more accurate dynamic predictions.

For detection of outliers we use the metric of identification rate. For identification of dynamic outliers for the humidity data the results are collected in Table 8. Results show that for method DPFFR, dynamic_FLR, and dynupdate, the mean rate of identification are 11%, 5%, and 51% respectively. The rate of identification is skewed to the right. The method dynupdate identifies more outliers than DPFFR and dynamic_FLR. Dynamic_FLR and DPFFR are overall in agreement about the identification of outliers for $r = 20$.

Figure 8 shows the rate of identification of dynamic outliers by different methods with cutoff point $r = 19$ and $r = 20$.



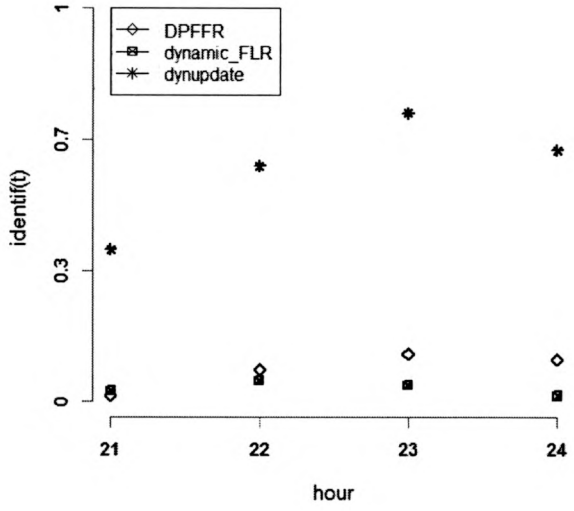
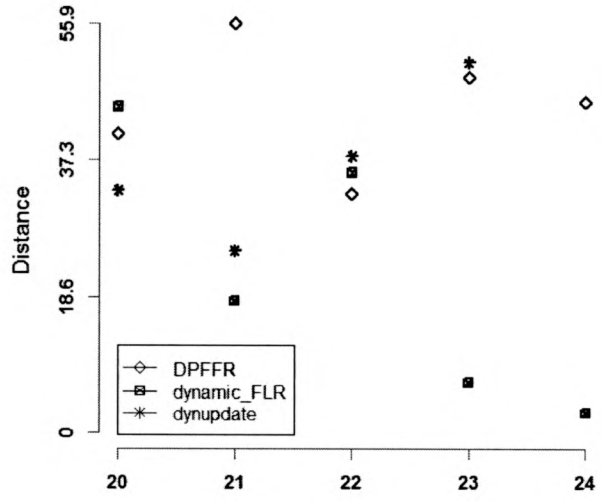


Figure 8: Hour-specific rate of identification of outliers by three methods. The panel in the first row shows the identification rate with $r = 19$ and the panel in the second row shows the identification rate with $r = 20$.

The above three curves show the patterns of identification rate at each time point for different methods when $r = 19$ and $r = 20$. Dynupdate has the largest rate of identification for outliers whereas dynamic_FLR and DPFFR has a lower rate of identification than dynupdate. Identification rate by DPFFR seems to be relatively the same as the time progresses. It can also be seen that at dynamic_FLR and DPFFR are rather close to each other in identification rate of dynamic outliers in this example. Towards the end of the day the DPFFR identified more dynamic outliers than earlier in the day.

If any dynamic outlier data point was detected, a mean distance was calculated. Table 8 gives the measure of distance of the outlier from the dynamic prediction intervals obtained by different methods in predicting humidity. Table 8 shows that with cutoff point 19, DPFFR has the lowest median value for the mean distance. When the cutoff point is 20

DPFFR has a similar median value for the mean distance as dynamic_FLR. As we increase the r value, the minimum, Q1, and median value of distance increases for DPFFR and this suggests that towards the end of the day humidity data that are dynamic outliers are more obvious. The highest the distance from the dynamic prediction interval, the more clear it is that humidity data is identified as dynamic outlier data. Figure 9 shows mean distance of dynamic outliers from the prediction intervals by different methods with cutoff point $r = 19$ and $r = 20$.



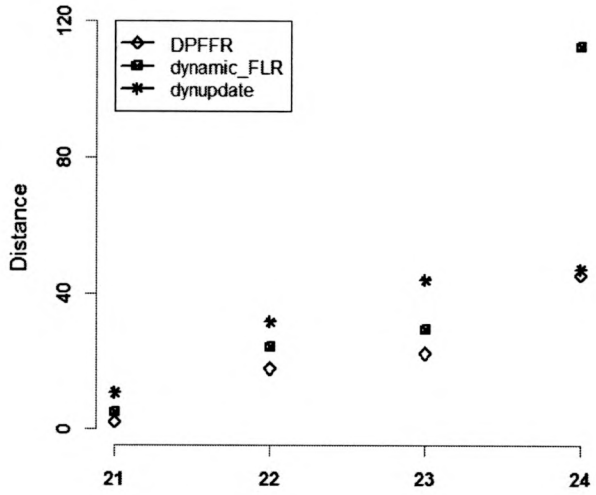


Figure 9: The panel in the first row is the graph of mean distance for three methods, with cutoff point $r = 19$. The panel in the second row is the graph of mean distance for three methods with cutoff point r =20.

The above three curves show the pattern for mean distance of dynamic outliers from prediction intervals results at each hour for three different dynamic prediction methods (DPFFR, dynamic_FLR, and dynupdate) when $r = 19$ and $r = 20$. With $r = 19$, we notice that in the beginning hour dynamic_FLR had the higher mean distance, but at the end DPFFR had the higher mean distance. With $r = 20$, dynupdate has a higher mean distance from prediction interval compared to other methods.

Furthermore, we obtained the width of the dynamic prediction interval by subject for each dynamic prediction method. Table 8 shows that as cutoff point increased the width increased slightly for DPFFR dynamic prediction intervals.

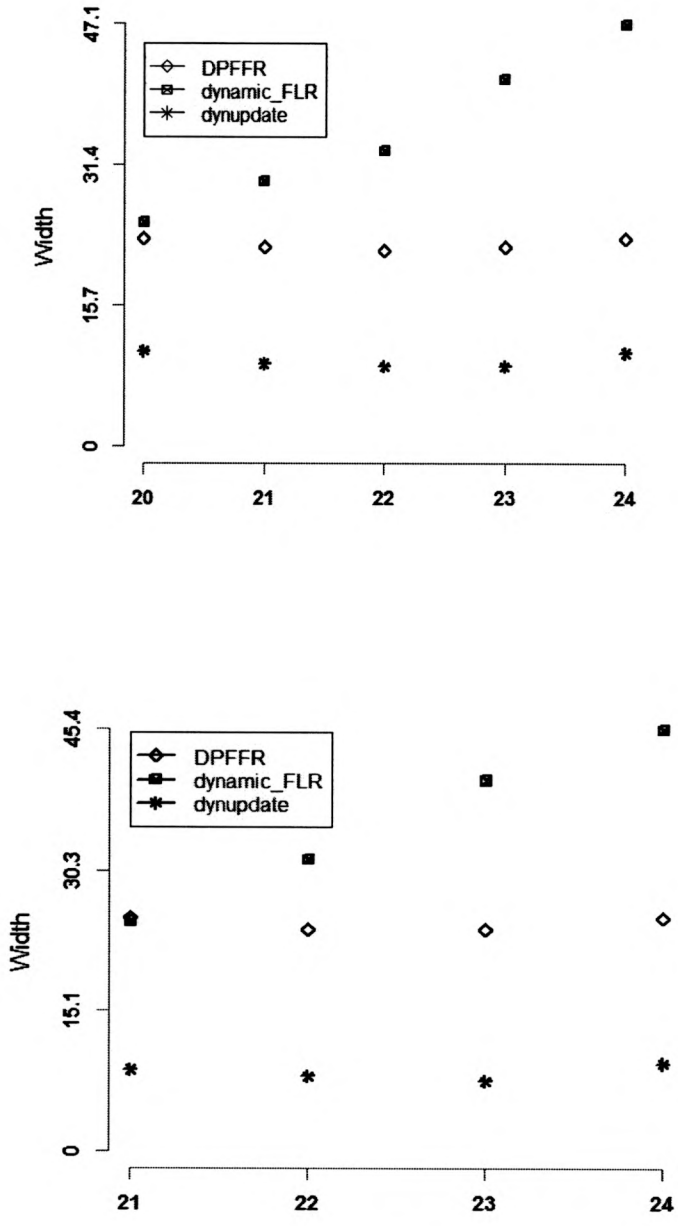


Figure 10 : Panel in the first row is the graph of mean width for three methods, with cutoff point $r = 19$. Panel in the second row is the graph of mean width for three methods with cutoff point r =20.

The above three curves in Figure 10 show the mean width for the dynamic prediction interval results at each hour for three different dynamic prediction methods (DPFFR, dynamic_FLR, and dynupdate) when $r = 19$ and $r = 20$. DPFFR has a lower prediction interval width compared to dynamic_FLR, and a wider width compared to dynupdate. Width by DPFFR seems to be relatively the same as the time progresses. It can also be seen that at initial points dynamic_FLR and DPFFR are close to each other, and towards the end of the time frame the differences in width are more pronounced.

Table 9 gives the overall numerical results computed as an average across all n samples and all points for $\tilde{t}$. We report IMPE, rate of identification of outliers, mean width, and computation time for different methods. CPU time was the computing time in seconds for the entire dataset consisting of $n = 75$ curves.

Table 9: IMPE, rate of identification, width and CPU time of different methods for identifying dynamic outliers in humidity data when $r = 19$ and $r = 20$.

r	Method	IMPE	Identification rate	Width of prediction interval	CPU time
$r = 19$	DPFFR	55.14	0.11	22.49	45.75
	Dynamic_FLR	63.29	0.05	34.82	207.93
	dynupdate	62.57	0.52	9.57	268.73
$r = 20$	DPFFR	49.44	0.08	24.50	42.42
	Dynamic_FLR	59.64	0.03	35.50	203.00
	dynupdate	55.30	0.59	8.43	270.19

The above table shows that for $r = 19$, DPFFR has a smaller IMPE than dynamic_FLR and dyupdate. Dynamic_FLR has the largest prediction interval width out of the three methods. CPU time shows that DPFFR has the fastest run time among three methods. Comparing the results for $r = 19$ and $r = 20$, we notice that IMPE with cutoff

point $r = 20$ is less than $r = 19$. The rate of identification was almost the same in for $r = 20$ and when $r = 19$. The average width of the prediction intervals and CPU time are similar for both cutoff points.

Graphical representation for prediction of dynamic outliers by different methods are included in Figure 11 and Figure 12.

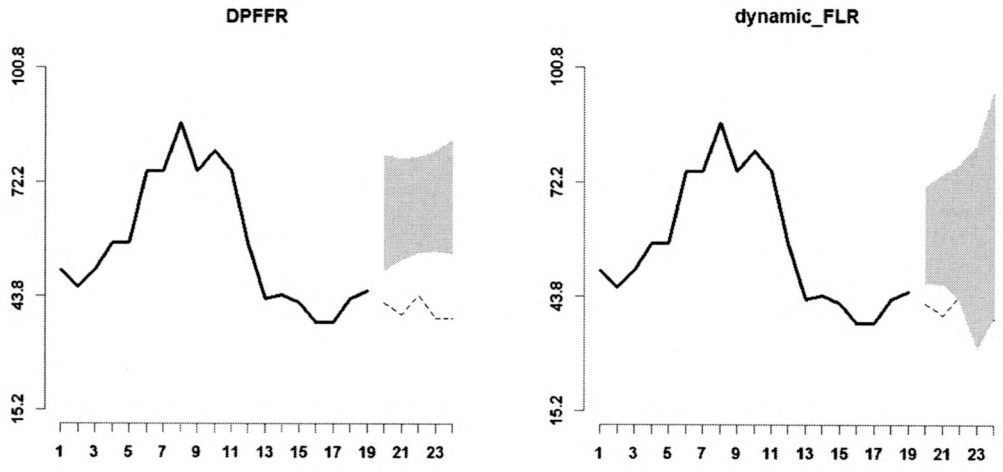


Figure 11: The left graph represents identification of dynamic outliers with cutoff point 19 corresponding to method DPFFR and the right graph is corresponding to dynamic_FLR for humidity data. The solid black line is the actual humidity up to the respective cutoff time point. The dashed line are the humidity data detected as dynamic outliers. The gray colored region represents the dynamic prediction intervals.

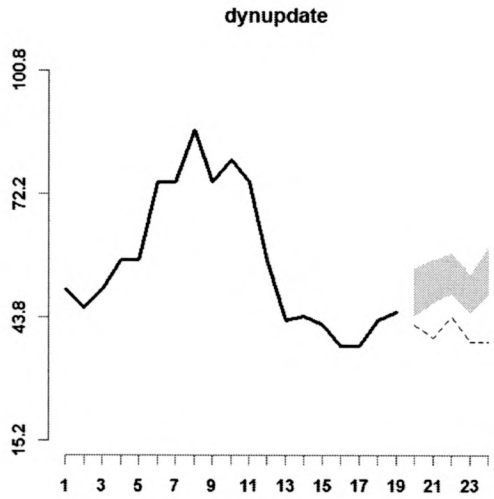


Figure 12: The graph represents dynamic identification of outliers with cutoff point 19 corresponding to method dynupdate for humidity data.

Figure 13 and Figure 14 indicate that DPFFR has narrower prediction interval than dynamic_FLR. DPFFR and dynupdate could identify humidity data as dynamic outliers at higher rate of identification than dynamic_FLR.

7 Discussion

Different methods have been studied for detecting dynamic outliers in the setting of functional data analysis. Comparisons have been made among several methods of dynamic prediction, prediction intervals, and identification of dynamic outliers when applied to several functional datasets. Results obtained from simulations and application to real datasets suggest that DPFFR works well and is among the preferred methods for detecting dynamic outliers. Almost in all simulation studies considered DPFFR can detect

dynamic outliers at a maximum or at a very high identification rate. Rate of identification of dynamic outliers increases when many curves and large size of outliers is observed.

References

Bunea F., Ivanescu A.E. and Wegkamp M.H. (2011). Adaptive inference for the mean of a Gaussian process in functional data. *Journal of Royal Statistical Society,* Series B, 73(4), 531-558.

Caballero, W., Giraldo, R. and Mateu, J. S. (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment,* 27(7), 1553- 1563.

Chiou, J. M. (2012). Dynamic functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics,* 6(4), 1588-1614.

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. The Annals of Applied Statistics, 3(1), 458-488.

Febrero-Bande, M., Galeano, P. and Gonzalez-Manteiga, W. (2007). A functional analysis of NOx level: location and scale estimation and outlier detection. *Computational Statistics.* 22, 3, 411-427.

Febrero-Bande, M., Galeano, P. and Gonzalez-Manteiga, W. (2008). Outlier detection in functional data by depth measures with application to identify abnormal NOx levels. *Environmetrics,* 19, 4, 331-345.

Febrero-Bande, M., Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software,* 51 (4), 1-28.

Goldberg, Y., Ritov, Y., and Mandelbaum, A. (2014). Predicting the continuation of a

with applications to call center data. *Journal of Statistical Planning and Inference,* 147, 53-65.

Goldsmith, J. and Scheipl, F. (2014). Estimator selection and combination in scalar-on-function regression. *Computational Statistics and Data Analysis,* 70, 262-372.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics,* 69, 41-51.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P.T. (2016). refund: Regression with functional data. Available at *https://CRAN.R-Project.org/package=refund.*

Hawkins, D. M. (1980). *Identification of outliers*. London-New York, Chapman and Hall.

Hyndman, R. J, and Shang, H. L. (2016). ftsa: Functional time series analysis. Available at *https://CRAN.R-Project.org/package=ftsa.*

Ignaccolo, R., Mateu, J. and Giraldo, R. (2013). Kriging with external Drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment,* 28(5), 1171-1186.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics*, 30(2), 539-568.

Martinez, T.J., Garcia, N.P.J., Alejano, L., Reyes. A.N. (2011). Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazardous Materials,* 186, 144-149.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4), 379-396.

Ramsay. J. O. and Dalzell C. J. (1991). Some Tools for Functional Data. *Journal of the Royal Statistical Society*, Series B, 53(3), 539-572.

Ramsay. J. O. and Silverman B. W. (2005). Functional Data Analysis. *Springer Science and Bussiness Media.* ISBN: 978-0-387-40080-8.

Ramsay. J.O., Hadley W., Spencer G., and Giles H. (2016). fda: Functional Data Analysis. R package version 2.4.4. *https://CRAN.R-project.org/package=fda.*

Sawant. P., Billor, N., and Shin. H. (2012). Functional outlier detection with robust functional principal component analysis. *Computational Statistics,* 27, 83-102.

Shang, H. L. (2015). Forecasting Intraday S&P 500 Index Returns: A Functional Time Series Approach. Available at *https//ssrn.com/abstract=2647233.*

Sorensen, H., Goldsmith, J. and Sangalli, L.M. (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine,* 32, 5222-5240.

UCI Machine Learning Repository. Available at http://archive.ics.uci.edu/ml/datasets/ *Bike+Sharing+Dataset.*

Yao, F., Muller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse Longitudinal data. *Journal of the American Statistical Association,* 100(470), 577-590.